Language Learning & Technology
ISSN 1094-3501

February 2020, Volume 24, Issue 1
pp. 169–186

ARTICLE

Synthetic voices in the foreign language context

Tiago Bione, Concordia University, Centre for the Study of Learning & Performance

Walcir Cardoso, Concordia University, Centre for the Study of Learning & Performance

Abstract

This study evaluated the voice of a modern English text-to-speech (TTS) system in an English as a foreign language (EFL) context in terms of its speech quality, ability to be understood by L2 users, and potential for focus on specific language forms. Twenty-nine Brazilian EFL learners listened to stories and sentences, produced by a TTS voice and a human voice, and rated them on a 6-point Likert scale according to holistic criteria for evaluating pronunciation: Comprehensibility, naturalness, and accuracy. In addition, they were asked to answer a set of comprehension questions (to assess understanding), to complete a dictation/transcription task to measure intelligibility, and to identify whether the target past -ed form was present or not in decontextualized sentences. Results indicate that the performance of both the TTS and human voices were perceived similarly in terms of comprehensibility, while ratings for naturalness were unfavorable for the synthesized voice. For text comprehension, dictation, and aural identification tasks, participants performed relatively similarly in response to both voices. These findings suggest that TTS systems have the potential to be used as pedagogical tools for L2 learning, particularly in EFL settings, where natural occurrence of the target language is limited or non-existent.

Keywords: *Text-to-speech, TTS, English as a foreign language, L2 pronunciation; L2 pedagogy*

Language(s) Learned in This Study: *English*

APA Citation: Bione, T., & Cardoso, W. (2020). Synthetic voices in the foreign language context. *Language Learning & Technology*, *24*(1), 169–186. https://doi.org/10125/44715

Introduction

Second language (L2) researchers and practitioners have explored the pedagogical capabilities of text-to-speech (TTS) synthesizers—speech synthesis applications that create spoken versions of written text—for their potential to enhance the acquisition of writing (Kirstein, 2006), vocabulary and reading (Proctor, Dalton, & Grisham, 2007), and pronunciation (Liakin, Cardoso, & Liakina, 2017; Qian, Chukharev-Hudilainen, & Levis, 2018; Soler-Urzua, 2011). Despite the positive evidence to support the use of TTS as a learning tool, the applications need to be formally evaluated for their potential to promote the conditions under which languages are acquired, particularly in an English as a foreign language (EFL) environment, as recommended by Cardoso, Smith, and Garcia Fuentes (2015).

This study evaluated a modern English TTS system in an EFL context in Brazil in terms of its speech quality, ability to be understood by L2 users, and potential for focus on specific language forms in comparison with a native English speaker, operationalized according to the following criteria for evaluating pronunciation: (1) text comprehension (i.e., users' ability to understand a text by answering comprehension questions); (2) intelligibility (the extent to which a message is actually understood), measured by a dictation-like (transcription) task; (3) users' ratings of holistic pronunciation features (comprehensibility, naturalness, and accuracy); and (4) users' ability to hear and identify a specific morpho-phonological feature (i.e., past *-ed*, produced as [t], [d], and [ɪd] depending on the preceding environment).

Literature Review

In her book *English Language and Technology*, Chapelle (2003) argues that, from both cognitive and social perspectives, Computer-Assisted Language Learning (CALL) tasks can offer L2 learners opportunities to receive enhanced input as well as interact with and produce the target language, all of which are recognized as essential for language acquisition. Previous research has investigated the effects of Computer Assisted Pronunciation Training (Thomson, 2012; 2018), Computer-Mediated Communication (Díez-Bedmar & Pérez-Paredes, 2012), Automatic Speech Recognition (Liakin, Cardoso, & Liakina, 2015), and Mobile Gaming (Grimshaw & Cardoso, 2018) on L2 learning. Among these options, one type of technology has stood out for its natural capacity to offer additional language input both inside and outside the classroom: Text-to-speech synthesizers.

Text-to-Speech Synthesis

Text-to-speech (TTS) is a type of speech synthesis application that creates a spoken (oral) version of textual input on personal computers or mobile devices. Handley (2009) explains that, in simple terms, "speech synthesis is the process of making the computer talk" (p. 906). This feature can be found on most modern computers, which now have the ability to "talk" via their built-in TTS features (e.g., Apple's Siri, Amazon's Alexa, and Google Translate).

The Benefits of Using TTS for Second/Foreign Language Acquisition

Some studies attest to the advantages of using TTS for developing different linguistic skills. To examine how TTS could support L2 English learners' writing processes, Kirstein (2006) analyzed data from six high school students. The data consisted of essays (written with and without TTS support), questionnaires, documents, interviews, and observations. Findings suggested that when participants used TTS, they wrote more drafts, spent more time on each draft, and detected more errors. Related studies have also found that TTS is useful for vocabulary acquisition and reading training, as its read-aloud functionality reduces the decoding demands of many challenging texts (Proctor et al., 2007).

Due to its aural nature, TTS seems to be particularly well-suited for pronunciation practice, as indicated earlier. Testing the effects of TTS on perception and production of /i/ and /ɪ/ (the vowels in "beat" and "bit", respectively), Soler-Urzua (2011), for instance, found that even though the TTS group outperformed the non-TTS and control groups, the overall improvement in the TTS group was not significantly different from the non-TTS group. Nevertheless, the author observed a trend showing improvements in perception and production by the TTS group, a pattern that was not observed for the other two groups.

In order to justify the pedagogical usefulness of TTS in the classroom, however, positive effects are not enough; prior to implementation, any SLA material must be evaluated for its pedagogical usefulness through well-established theoretical frameworks to produce reliable and comparable results (Jamieson & Chapelle, 2010). Hence, TTS needs to be thoroughly examined in light of relevant theory and research in SLA before being promoted as a pedagogical tool.

TTS Evaluation: Speech Quality

An initial step toward evaluating synthetic speech is to assess how it differs from natural speech. In other words, how does the quality of modern synthetic voices compare to human voices? To analyze TTS speech quality, researchers have drawn on previous studies of listeners' reactions to non-native speech. For instance, to evaluate L2 speakers' pronunciation in general, Derwing and Munro (2005) proposed three dimensions of oral speech: (1) comprehensibility, or how difficult it is to understand an utterance; (2) intelligibility, or the extent to which a message is actually understood by an interlocutor or group of listeners; and (3) accentedness, or how much an L2 accent differs from the L1, including the variations in accents that characterize native or fluent speech.

In the context of synthetic voices, which are produced by software applications that try to emulate specific "native" language varieties or accents (e.g., North American English, British English), the concept of

accentedness needs to be modified to capture how much the synthesized speech deviates from that of a human. In this study, based on insights from Cardoso et al. (2015), we decompose the accentedness construct into (1) naturalness, or the extent to which the TTS voice deviates from that of a human (see also Dall, Yamagishi, & King, 2014); and (2) accuracy, or the extent to which the TTS voice accurately reproduces human speech.

There have been a few evaluations of TTS systems and their voices over the past two decades. The favored method has been to judge TTS and human speech samples under the set of categories mentioned above. For example, in a study by Nusbaum, Francis, and Henly (1995), the authors compared TTS-produced voices in English to their human counterparts for naturalness in both segmental and suprasegmental features. In their first experiment, they instructed native English-speaking participants to evaluate utterances of the segments /a/, /i/, and /u/ using a naturalness scale to measure the probability of a sound to be considered natural. Results differed between vowel categories, as TTS was perceived to be more natural than human voices for /a/, less natural for /i/, and equally natural for /u/. In a second experiment, L1 English participants evaluated prosody at the word level, also using a naturalness scale. The researchers manipulated the input to isolate prosody by removing all the segmental information from the stimuli. Therefore, participants were only able to listen to rhythmic word patterns produced by TTS and human voices. Their findings showed that even with the intelligibility variable removed, participants would still judge human voices to be more natural than TTS. Stevens, Lees, Vonwiller, and Burnham (2005) echoed these results when they found that their native English-speaking participants rated TTS sentences to be less natural than human-produced sentences. Other studies, however, have found more positive results for TTS voices regarding naturalness. For instance, Kang, Kashigawi, Treviranus, and Kaburagi (2008) asked Japanese-speaking participants to rate English TTS and human input at word and sentence levels. They found that TTS voice was perceived to be as natural as human production, at least at the word level. These results are partially substantiated by Stern, Mullennix and Yaroslavsky's (2006) findings, as they observed TTS messages to be perceived as favorably as those produced by humans.

Other TTS evaluations have focused on cognitive factors in synthetic voice comprehension. Delogu, Conte, and Sementina (1998), for instance, designed an experiment in which participants listened to short paragraphs in synthesized Italian, then completed a multiple-choice comprehension test designed to objectively evaluate the voices in terms of intelligibility. Their results demonstrated that, in general, comprehending synthetic voices is more demanding, as response duration is higher and the degree of text comprehension is lower. Still, the authors indicated that the difficulty level decreased as the participants had more exposure to synthetic voices. In another study that focused on measuring intelligibility using a French TTS system, Bailly (2003) noticed that participants performed better in shadowing tasks when they had human voice input instead of TTS-produced input. Interestingly, in a more recent study, Kang et al. (2008) found no significant differences between their participants' ability to understand human and TTS speech in text comprehension.

Based on the handful of studies available, it seems clear that previous research has yielded mixed results regarding the quality of TTS systems compared to the human voice. One reason for this discrepancy is the use of inconsistent or incomparable methods. For example, rather than taking a comprehensive, holistic view on the assessment of TTS-produced voice quality, previous studies have used different criteria in their evaluations: While some focused exclusively on users' perceptions regarding the synthetic voice's naturalness, (e.g. Nusbaum et al., 1995; Stevens et al., 2005), others have included only comprehension measures (e.g. Bailly, 2003; Delogu et al., 1998). In addition, most studies have used native speakers as TTS evaluators, which may have impacted their results and, therefore, they cannot be generalizable to L2 speakers—the target population of the current study. Furthermore, those investigations are relatively dated, with the most recent being from 2009. Text-to-speech synthesis has evolved considerably over the past two decades, particularly since the advent of voice-based personal assistants found in GPS systems, smartphones (Siri, Cortana), and speaking assistants (Amazon Echo, Google Home). Finally, previous studies have not investigated TTS's potential for focus on specific language forms, which is a crucial element in evaluating the effectiveness of any tool for L2 pedagogy.

One exception to this is a recent study by Cardoso et al. (2015), in which an evaluation of an up-to-date English TTS system's speech quality and potential to draw students' attention to linguistic forms was performed. Moving beyond previous studies, a new layer was added to evaluate TTS in terms of its potential to allow learners to focus on language. The task targeted the aural identification of English past tense *-ed* allomorphy: [t], [d], and [ɪd], as found in inflected past forms such as "walk[t]", "drag[d]" and "add[ɪd]", respectively. Fifty-six university-level students in an English-speaking university in Montreal, Canada performed a series of tasks to evaluate a current TTS system, in which they heard utterances alternately produced by TTS and human voices. Both native and second language speakers participated in this study. Results showed that the samples produced by the TTS system were rated significantly lower than the human-produced samples for all four categories of speech quality (comprehensibility, naturalness, accuracy, and intelligibility). However, excluding naturalness, TTS rating was still considered high (above 80% for comprehensibility, accuracy, and intelligibility). Regarding the potential to focus on a linguistic form, the TTS and human-produced samples had similar results, indicating that regardless of the source of delivery (human or TTS), participants were equally able to perceive the target past *-ed* allomorph (/t/, /d/, or /Id/) in decontextualized, atemporal phrases. The implication of this finding is that modern TTS systems are ready for use in the L2 learning context, particularly as a supplemental source of language input. In their conclusion, the authors suggested directions for future research by calling for studies involving foreign language contexts, particularly those in which opportunities for naturally-occurring English input are scarce or non-existent. Thus, the goal of this quasi-replication study is to address this recommendation in an English as a foreign language setting.

Differences in Learning Contexts: ESL Versus EFL

It is attested in the EFL literature that, in the EFL context, students may have low exposure to the target language, both within and outside of the classroom environment (Collins & Muñoz, 2016). Foreign language class time is often limited to few hours a week, which is not enough to provide students with the amount of input and practice necessary for mastering L2 skills. In Brazil, for instance, the quantity of L2 English exposure is reduced to two hours of instruction per week in the public-school system (British Council Brasil, 2015). Ortega (2013) estimates that whereas students in *second* language contexts may accrue 7,000 hours of L2 exposure in five years of contact with the target language (in a conservative projection of 4 hours of exposure a day), *EFL* students, on the other hand, may have as little as 540 hours of L2 exposure from instruction only in the same period (i.e., less than 10% of what is observed in Ortega's conservative estimates for *second* language contexts).

Although EFL students can access the internet and thus have at their disposal a large amount of English oral input from native speakers of a wide range of dialects, as well as L2 speakers, they cannot easily self-select speech stimuli from naturally produced tokens (e.g., they cannot isolate specific phonetic features or control the type and speed of the target speech). An interesting pedagogical affordance of TTS is that it can be used for remedial purposes (e.g., if a Francophone student is having problems discriminating English /h/, the teacher could give that student targeted exercises that focus on that sound —for instance, minimal pairs such as air-hair, hear-ear, etc.). Therefore, by having less focused exposure to their target language outside the classroom, EFL students tend to greatly rely on their teachers for their L2 input (Tanaka, 2009), which can create a teacher-centered environment that is not ideal for learning (Chapelle, 2001). This environment is particularly unproductive for pronunciation training, as exposure to L2 phonology is limited to one teacher who uses only one variety of English or one accent. It also goes counter to the best pedagogical practices, which recommend a learning environment in which the aural input is highly variable (see Thomson, 2018, for the positive effects of High Variability Pronunciation Training (HVPT) on L2 learning). One of the affordances of TTS is its ability to create a learning environment in which different voices (e.g., based on gender, age, accent, pitch) can be selected and manipulated, which can consequently promote the implementation of HVPT principles in L2 pedagogy.

Aware of these limitations, one may conclude that considerable dissimilarities in language exposure and learning settings may create distinctive demands from and for ESL and EFL students. Thus, it is

hypothesized that a change in learning environment (from second to foreign) may positively affect learners' perceptions of and attitudes towards TTS-produced speech, as EFL students may perceive synthetic voices as a useful source of additional input, given the often-limited exposure to the target language in their learning environment.

The study

The objective of this study is to evaluate the quality of a typical TTS voice in comparison with that of a human, and consequently examine its pedagogical potential for use in an EFL setting, following Cardoso et al.'s (2015) recommendation. This study is guided by the following research question: What is the quality of speech produced by a TTS system in comparison with that of a human, based on the following six assessment measures:

1. text comprehension (one's ability to understand a short anecdote—a type of intelligibility measure that is more cognitively effortful than speech transcription)
2. intelligibility (the extent to which a message is actually understood by an interlocutor or group of listeners)
3. comprehensibility (one's perception of how easy it is to understand a message)
4. naturalness (the extent to which the TTS voice deviates from that of a human)
5. pronunciation accuracy (the extent to which the TTS accurately reproduces human speech)
6. form identification (the participant's ability to identify linguistic forms in speech: The identification of past -ed forms)

Methodology

Participants

Twenty-nine Brazilian EFL adult learners (Male = 9, Female = 20) participated in the study that took place in Recife (Brazil). Their ages ranged from 18 to 33 (Mean = 23.6, *SD* = 4.9), and all spoke Portuguese as their first language (L1). Their proficiency level was intermediate, established by four criteria: (1) placement at their language institution; (2) the call for participants (which emphasized the target language proficiency); (3) their self-assessment in a background questionnaire; and finally, (4) the researcher's overall perception of their skills (e.g., if they could not follow instructions in English or could not understand the written materials, the participants were not included in this study). This proficiency group was targeted not only because intermediate learners are capable of judging naturalness (Major, 2007), but also because they represent the typical clientele that could benefit from TTS (i.e., those who have not mastered the L2). In addition, previous research has already examined native and advanced ESL learners (Cardoso et al., 2015).

Design

This study considered two independent variables, TTS and human voice, and measured their effects on three variables: (a) intelligibility (in text comprehension and dictation/ transcription, as will be described below), (b) learners' ratings on holistic pronunciation measures (comprehensibility, naturalness, and accuracy), and (c) their ability to aurally identify the morphophonemics of a grammatical form (past -ed).

The data were collected in one individual session wherein each participant completed a set of tasks designed to assess the abovementioned criteria. For intelligibility, participants listened to two short stories and answered six multiple-choice questions covering each story's main points (e.g., Why was the bird so special?). In addition, participants completed a dictation task during which they were asked to transcribe TTS- and human-based utterances on an answer sheet, as suggested by Derwing and Munro (2005).

In order to evaluate pronunciation holistically, participants rated the quality of the speech that they heard

with the two short stories described above based on three categories: Comprehensibility ("How easy was the voice to understand"), naturalness ("How natural was the voice?"), and pronunciation accuracy ("How correct was the voice?"), using a 6-point Likert scale ranging from "very difficult" to "very easy", "very unnatural" to "very natural" and "very incorrect" to "very correct," respectively (note that these statements were pilot-tested among a group of L2 learners for ease of understanding and accuracy). Participants also rated 12 short sentences (e.g., "The boy watched the clock ticking on the wall"). The rationale for the inclusion of these short sentences was that they could yield different results due to the low cognitive load required for their processing, as the participants needed to concentrate solely on speech quality, not understanding. Short stories, on the other hand, contain more complex structures and may require more cognitive effort, which may impact intelligibility and participant ratings.

Finally, for the ability to focus on a linguistic form, as in Cardoso et al. (2015), participants performed an aural identification task for 16 sentences in which they judged whether the target feature (past tense *-ed*) appeared in the oral input they heard (sentences in the present tense (n=4) were included as distractors). Participants had to decide if the action took place in the past (e.g., "I called my mother") or not (e.g., "I visit my cousin Sam") and check the corresponding form on the answer sheet.

Stimuli

For all tasks, participants listened to speech samples that randomly alternated between TTS and human voices. The TTS voice was Julie (by NeoSpeech; http://neospeech.com), a female North American speaker whose voice was used for the synthesis of the target texts and sentences (see material description below). Human speech was produced by a female native speaker of the same North American dialect, with speech properties (e.g., female, midrange pitch) similar to those of the synthesized voice. She recorded the same text and sentences as Julie and was instructed, whenever possible, to match Julie's speed so that each recording pair would have roughly the same duration and tempo. Audio samples of two sets of recordings are available here, where two short stories (used for pronunciation ratings and intelligibility assessment) and two sentences (used for pronunciation ratings) are provided.

Both human and TTS samples were recorded in WAV audio format (Stereo, 16bit, 44.1KHz) and later converted into Mono (to comply with the monophonic nature of the human voice and to save space). These samples were then embedded in Microsoft PowerPoint slides for presentation to the participants.

Materials

Both the stories and sentences were adapted from materials produced by the ALERT research project (e.g., Collins, Trofimovich, White, Cardoso, & Horst, 2009). Each short story had approximately 230 words and lasted for approximately the same amount of time, regardless of voice type: 1min:43sec and 1min:22sec for Julie's output, and 1min:44sec and 1min:20sec for the human-recorded text. The text comprehension task for each short story consisted of six multiple choice questions, each with one correct and three incorrect responses to choose from. The questions were divided into two types: Specific (e.g., Why did the woman go to the store?) and general (e.g., What do you think happened after?). Participants could score from 0 (if no correct answer was selected) to 6 points (if all six answers were correctly selected) on each test involving text comprehension.

In addition to short stories, participants were exposed to 38 short sentences in total for the three remaining tasks (mean word count for each sentence was 9 words, $SD = 3.7$. All words, including content and function words, were considered in this computation), corresponding to 2–3 seconds of speech for each. For the intelligibility assessment of sentences, 10 utterances were generated for a task similar to a dictation, where participants heard each sentence only once, after which they were asked to transcribe what they heard on an answer sheet. It was assumed that sentences that were more intelligible would yield a higher percentage of correctly transcribed words. Students' orthographic inaccuracies were ignored, as the task was not intended to measure writing skills, but rather the extent to which participants could hear and comprehend English utterances. Participants could score from 0% to 100% on each sentence, where 0 corresponded to no intelligibility at all and 100% represented the highest intelligibility level.

For the holistic assessment of pronunciation in terms of comprehensibility, naturalness, and accuracy, 12 sentences were designed to match the vocabulary and morphosyntactic knowledge of intermediate-level learners so that the participants could focus exclusively on these three impressionistic measures. In addition, the target sentences were constructed without any references to contextual cues such as "last week" or "every day."

Finally, the linguistic feature identification material (past *-ed*) consisted of 16 sentences carefully designed to avoid any lexical cues that could help participants to identify the tense (e.g., words such as yesterday, usually, etc.) without using morpho-phonological processing. This way, participants' judgments were taken based solely on their aural perception. After listening to each sentence once, participants were asked to decide whether the action took place in the present or past. Table 1 shows the distribution of present- (included as distractors) and past-tense sentences in the stimuli, as well as the allomorphic distribution among the past sentences.

Table 1. *Distribution of Present/Past Sentences and Allomorphy*

Tense	**Total**	**/d/**	**/t/**	**/ɪd/**
Present (Non-past)	4	-	-	-
Past	12	3	4	5

The instruments used in this study, the measures they are designed to test, the tasks in which they were included, and how the data that they elicited were analyzed are summarized in Table 2.

Table 2. *Summary of Instruments, Measures, Tasks, and Analysis*

Instrument	**Measure(s)**	**Tasks**	**Analysis**
Text comprehension (short stories)	Intelligibility	Comprehension test (n=2; 12 items)	Average of correct answers
Dictation (sentences)	Intelligibility	Sentence transcription (n=10)	Percentage of transcribed words
Learners' ratings	Comprehensibility Naturalness Pronunciation Accuracy	Stories (n=2) and sentence (n=12) ratings	Average of ratings in a 6-point Likert scale
Aural identification of past tense -ed	Ability to identify a grammatical form	Tense identification (n=12 target items)	Percentage of correct identification

Procedures

To complete all tasks, participants had approximately one hour in one individual session. Before the session started, they were asked to read and sign a consent form, after which they received a brief description of the project and rating categories. However, it was not disclosed to the participants that they were going to listen to different voice types (including synthesized voices) among the samples. Participants then proceeded with the Microsoft PowerPoint presentation to initiate the study. They listened to the target stimuli using headsets (Microsoft Lifechat LX-3000), wrote their answers, rated voices, and completed the dictation task on a printed answer sheet as they advanced task by task in the presentation.

The material presented to participants was organized in two randomized sequences (Sequence A and Sequence B) in a way that both sequences contained the same target sentences or texts but were produced by different voice sources. For example, participants who received Sequence A heard the same sentences as participants in Sequence B; however, all the sentences produced by TTS in Sequence A were recorded by human voice in Sequence B, and vice-versa.

Results

Participants' judgments of the stories and sentences (to measure comprehensibility, naturalness and accuracy), text comprehension results, percentage of correct words in their dictation task (both to measure intelligibility), and their accuracy on identifying regular past (to measure TTS's ability to provide noticeable input) were tallied, and means of matched pairs were compared. Parametric statistics were used for data sets that meet the normality assumptions (namely data from short stories' text comprehension and ratings); for every other set, non-parametric tests were carried out. Paired sample t-test and Wilcoxon Signed-Rank tests were used respectively, with an alpha level of .05 for the determination of statistical significance. An adjusted alpha of .004 was calculated using a False Detection Rate (FDR) post-hoc method to avoid false positive errors. As the Bonferroni adjustment may be too conservative when the number of comparisons is high (this study includes nine comparisons), which may lead to false negative errors, an FDR was deemed more suited for this analysis (see Herrington, 2002).

Intelligibility

Intelligibility was measured at two cognitive levels: A text comprehension test for short stories (complex cognitive level) whose scores could vary from 0 to 6 on each story depending on how many questions were correctly answered, and sentence transcription (simple cognitive level), where participants could transcribe between 0% to 100% of each sentence depending on the number of words correctly transcribed. The details for each analysis are described below.

Short Stories (Text Comprehension)

A Paired sample t-test was conducted to compare how intelligible TTS- and human-narrated short stories were. There was no significant difference in the scores for TTS (M = 4.57, SD = .81) and human (M = 4.74, SD = .75); t(1) = -4.25, p = .147. Figure 1 illustrates the results for each story. These results suggest that the type of voice input that the participants received to complete the listening comprehension task had no impact on intelligibility for either story.

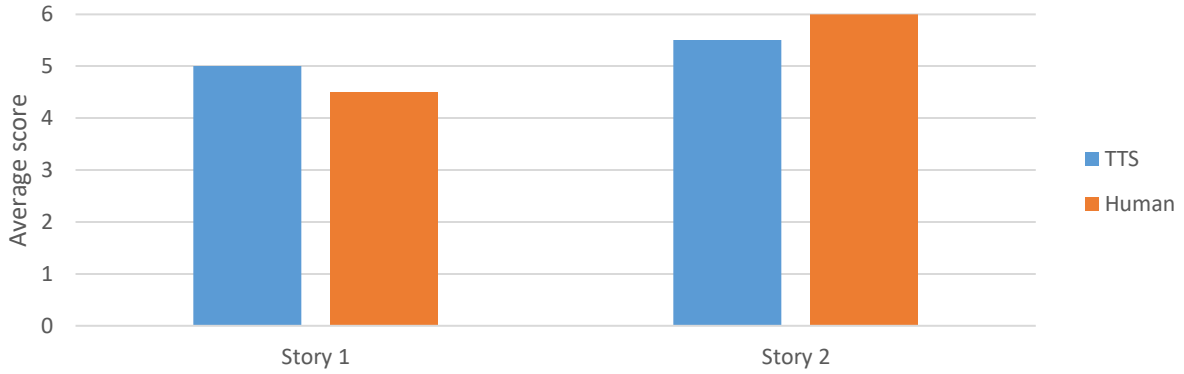

Figure 1. Short stories' score average.

Sentences (Dictation Task)

A Wilcoxon Signed-Rank test was conducted to compare intelligibility in sentences produced by either TTS or human voice. There was no significant difference in the scores for TTS (Mdn = 59.65%) and human samples (Mdn = 55.05%); Z = -.153, p = .878. As roughly 60% of all words within the sentences were transcribed, regardless of their source, these results show that the type of voice did not affect intelligibility at simple cognitive levels such as listening to sentences.

Figure 2 shows the distribution of correctly transcribed words for each sentence pair (spoken by TTS or a human) by all participants (where 1 = A four-year-old boy sat in the doctor's waiting room with his mother; 2 = He saw a pregnant woman on the other side of the room; 3 = Is the baby in your stomach?; 4 = If he is

such a good baby, then why did you eat him?; 5 Last Christmas, Jimmy received the best present: It was a parrot; 6 = Jimmy heard the parrot say some very bad words; 7 = Jimmy was so frustrated that he decided to punish the bird; 8 = He carried his parrot into the kitchen and put it in the freezer; 9 = He did not know why the parrot stopped saying bad words after only a few minutes in the freezer; and 10 = May I ask what the chicken did wrong?). To examine whether there were differences in error types among the two voice sources, we conducted an error analysis of the types of errors that the participants made in their transcriptions. We found that voice source had no effect on the participants' transcriptions; for instance, the transcriptions for both the TTS and human voices similarly reflected the intended text. As expected, due to cognitive effort, shorter sentences such as "Is the baby in your stomach?" (n=6) were more accurately transcribed (intelligible) than longer ones such as "He did not know why the parrot stopped saying bad words after only a few minutes in the freezer" (n=19), regardless of voice source.

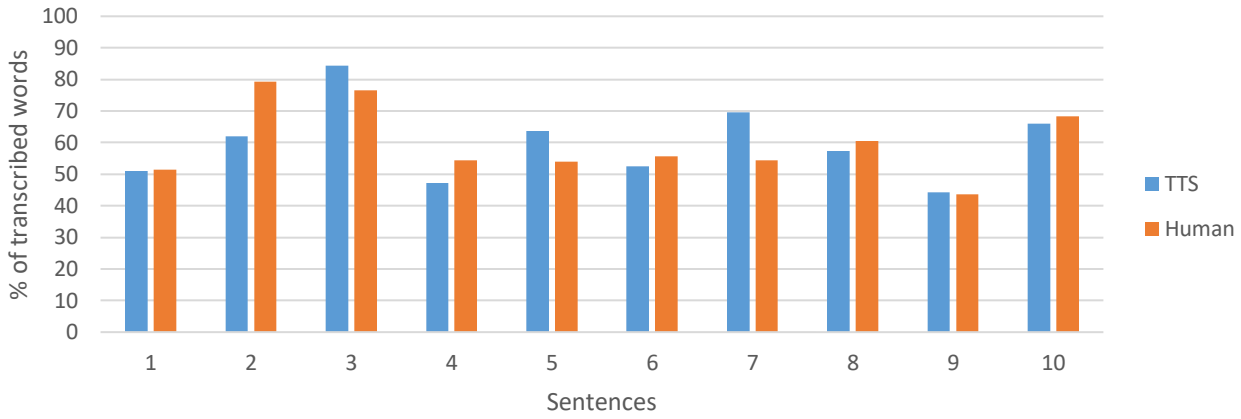

Figure 2. Accurately transcribed words in sentences (%).

Users' Ratings: Comprehensibility, Naturalness, and Accuracy

For each measure under users' ratings, paired sample t-test (for short stories' ratings) or Wilcoxon Signed-Rank tests (for sentence ratings) were conducted. Statistical test results are reported below.

Short Stories

Considering an adjusted alpha of .004, paired sample t-tests yielded no significant difference in ratings for any pronunciation measure included in the study, as shown in Table 3. These results indicate that, when listening to short stories, participants did not find substantial dissimilarities between TTS and human-produced samples regarding comprehensibility, naturalness and accuracy.

Table 3. *Short Story Holistic Ratings*

	TTS		Human			
Measure	*M*/6	*SD*	*M*/6	*SD*	***t***	***p***
Comprehensibility	4.42	0.02	4.92	0.30	-2.59	0.235
Naturalness	3.12	0.74	4.58	0.41	-6.35	0.099
Accuracy	5.04	0.15	5.31	0.13	-27.00	0.024

Sentences

Based on Wilcoxon Signed-Rank tests, human samples were considered significantly more natural and more accurate than TTS samples. On the other hand, no significant difference was found for comprehensibility, as illustrated in Table 4.

For a more detailed illustration of the results presented in Table 4, Figure 3, Figure 4 and Figure 5 provide

the distribution of user ratings for each of the 12 sentences used in the respective test (where 1 = He placed the glasses on his nose and looked up; 2 = When he arrived, he saw that the front door was open; 3 = She quickly opened the box and found the pictures and the letter; 4 = I looked for your picture, but I can't remember which girl you are; 5 = He stood up and walked to the chair where she was sitting; 6 = The boy watched the clock ticking on the wall; 7 = He talked to his mother very politely and said very nice things; 8 = His mother and father explained that bad words were not polite; 9 = The boy stepped back from the fence and rolled up his pants; 10 = The girl put her hand into her pocket and pulled out a handful of change; 11 = The teacher talked for twenty minutes about school and being good students; and 12 = My teacher asked me to please sit down).

Table 4. *Sentence Holistic Ratings*

Measure	TTS *M*/6	Human *M*/6	*Z*	*p*
Comprehensibility	5.06	5.10	-0.628	0.530
Naturalness	3.45	5.13	-3.06	0.002*
Accuracy	4.93	5.10	-2.85	0.004*

**p< .004*

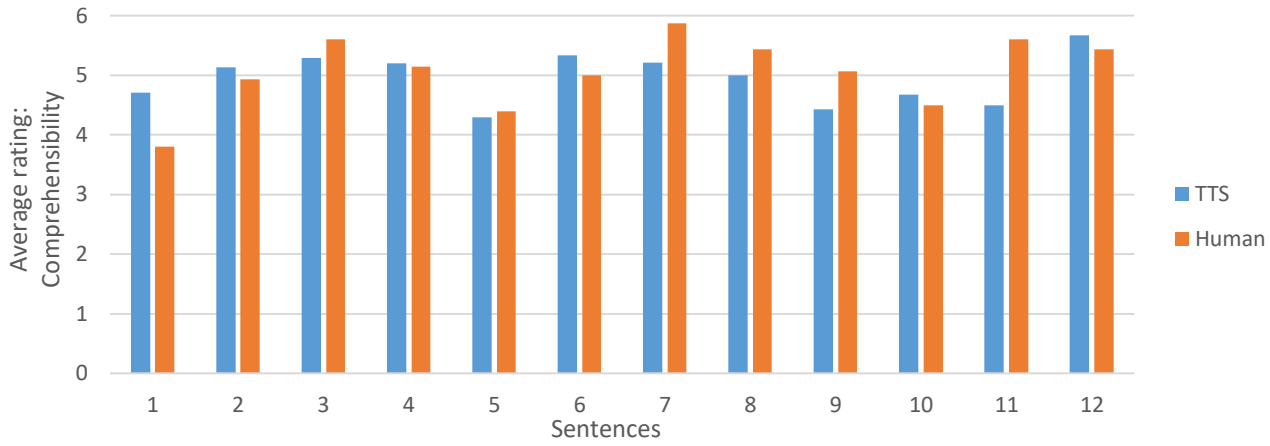

Figure 3. Comprehensibility ratings for sentences.

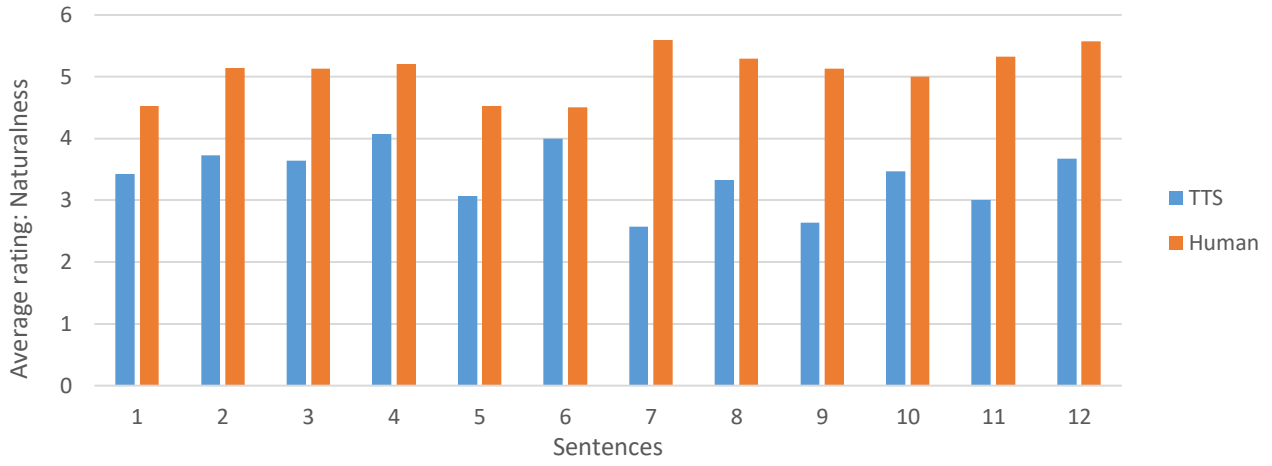

Figure 4. Naturalness ratings for sentences.

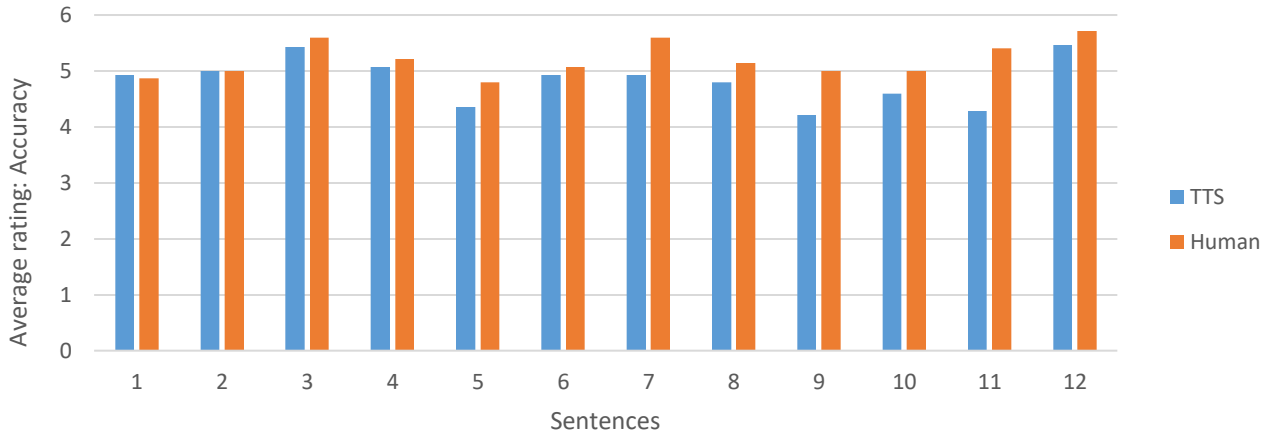

Figure 5. Accuracy ratings for sentences.

Combining the results obtained for short stories and sentence listening, they indicate that text complexity may affect participants' ratings, since TTS was rated as natural and accurate as the human voice in the presence of cognitively complex input (short stories), but received significantly lower ratings for those two measures when cognitive complexity decreased (sentences). Conversely, comprehensibility seems unaffected by text complexity, as TTS and human voice were equally comprehensible for participants at both simple (sentence rating) and complex input levels (story rating).

Aural Identification of a Linguistic Form: Past *-ed*

A Wilcoxon Signed-Rank test showed no significant difference in accuracy in identifying past -ed verbs between voice types. In other words, participants were equally able to recognize if a sentence was set in the simple past for both TTS (Mdn = 76%) and human samples (Mdn = 85%); Z = -1.735, p = .083. Figure 6 displays the percentage of correct identification by voice type for each past tense sentence (excluding the distractors in present tense form; where 1 = I called my mother; 2 = I talked with Jeff in the hallway; 3 = I grilled the hamburgers; 4 = I corrected my math homework; 5 = I jumped in the freezing lake in winter; 6 = I invited him to dinner; 7 = I opened the door for her; 8 = I fixed the problems around the house; 9 = I hated the movie; 10 = I danced to the music; 11 = I waited two hours for my friend; and 12 = I painted some pictures).

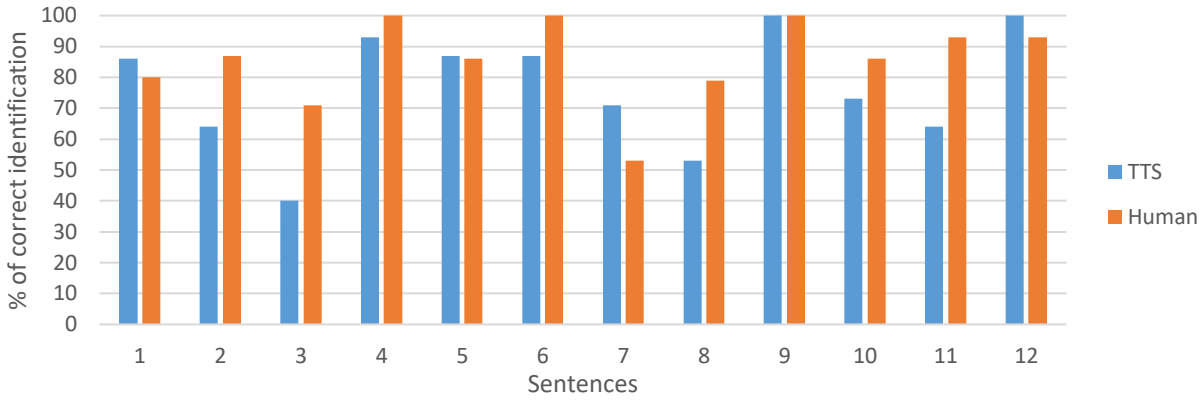

Figure 6. Aural identification of past *-ed* forms in sentences.

Participants behaved similarly with the distractors (present tense) since the data did not show a noticeable difference between voice types. For a comprehensive distribution of results regarding the participants' ability to identify both past and present forms in the target voices as well as the representation of past tense allomorphy in the target sentences, see Figure 7.

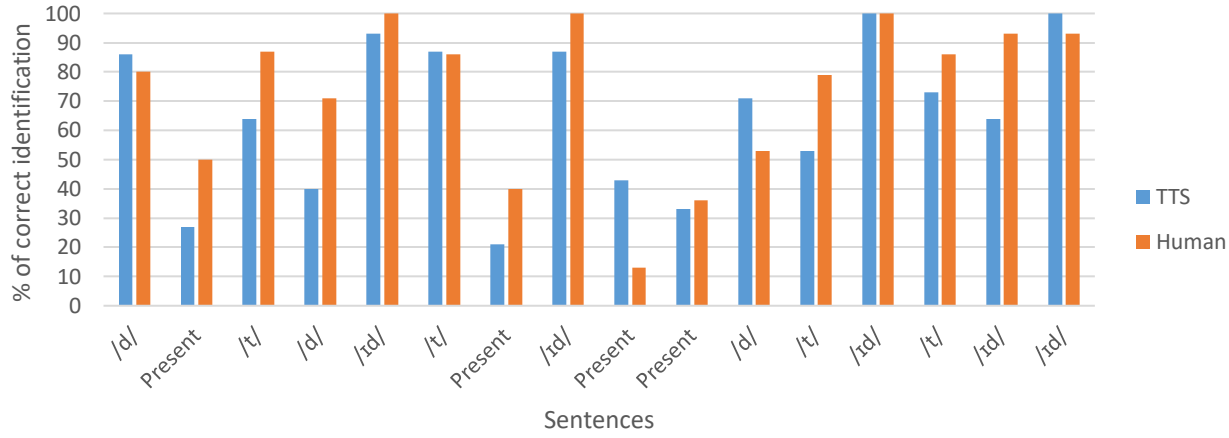

Figure 7. Aural identification of past forms and distractors (present) by *-ed* allomorphy: sentences.

Discussion

This study evaluated the voice quality of a TTS system in comparison with a human voice, and consequently examined its pedagogical potential for use in an English as a foreign language setting. The following research question was addressed: What is the quality of speech produced by a TTS system in comparison

with that of a human? The answer to this question was based on six assessment measures of aural abilities: text comprehension (one's ability to understand a short anecdote), intelligibility (the extent to which a message is actually understood by interlocutors or listeners), comprehensibility (one's perception of how easy it is to understand a message), naturalness (the extent to which the TTS voice deviates from that of a human), pronunciation accuracy (the extent to which the TTS accurately reproduces human speech), and opportunities for grammatical feature identification (operationalized as one's ability to identify regular past tense in sentences). These measures encompass three general aspects for assessing TTS-generated speech (Cardoso et al., 2015, influenced by Derwing & Munro, 2005): Intelligibility (at two distinct cognitive levels: Complex short stories and simple short sentences), users' holistic ratings (including comprehensibility, naturalness and pronunciation accuracy), and their ability to focus on a linguistic form (past *-ed*).

Analysis of the data collected in the study showed that EFL learners rated or performed similarly, regardless of the input source, except for the naturalness and accuracy measures at the sentence level only (not in longer narratives). Overall, these results correspond to what Kang et al. (2008) found in their research involving non-native English speakers, wherein they concluded that L2 learners do not recognize a difference between synthetic and human voices. A discussion of the results obtained for each feature under investigation is provided below.

Intelligibility: Text Comprehension and Dictation Task

Previous studies have most commonly reported that TTS presents low intelligibility when compared to natural speech. For instance, Delogu et al. (1998) concluded that the user's cognitive load is heavier in synthetic voices because listening to TTS is a more demanding task than listening to humans, possibly due to the unexpected pauses and/or other prosodic limitations observed in synthesized voices. Bailly (2003) presented similar results, as his participants performed better in shadowing tasks involving human voices than those using artificial voice samples. Contrary to previous studies where TTS scored lower than human voice, the current research revealed that both voice sources were equally intelligible. This contrast may be due to two factors: The new advances in TTS technology and the participants' increased exposure to electronic voices, as will be discussed next.

Elaborating upon the poor results previously reported for TTS, Bailly (2003) suggests that they were mainly due to the inappropriate prosody generated by the technology available at the time. It is outside the scope of this study to compare current and previous versions of TTS applications, but if we consider that almost 15 years have passed since Bailly's publication, we may comfortably assume that speech technology has advanced considerably. As indicated by Handley (2009) six years later, current TTS systems have not yet reached an optimal development stage at the prosodic level; however, the data presented in this study show that they have at least evolved to the point where their voice quality does not affect intelligibility.

Regarding the hypothesis that an increase in exposure to electronic voices may lead to a higher acceptability rating, Delogu et al. (1998) noticed that intelligibility increased when participants became more acquainted with synthetic voices. If Delogu et al.'s remarks about a positive correlation between exposure to electronic voices and intelligibility are accurate, then increasing access to these types of voices may explain why this study found no difference in intelligibility between synthetic and human voices. Most current computers and mobile devices offer built-in text-to-speech capabilities (e.g., via GPS systems, smartphones, and personal assistants such as Google Home, Amazon Echo and Apple HomePod), which increase users' reliance on artificial speech. In addition, it is virtually impossible to contact a service provider without first interacting with an electronic voice that guides customers through menus before a human agent is reached. Although the current study did not measure participants' previous experience with these types of synthetic voices, we can ascertain that, due to their age (young and educated adults) and the ubiquitous use of synthetic voices in phone-based customer service, they are regularly exposed to TTS-generated voices.

Learners' Ratings of Holistic Pronunciation Measures: Comprehensibility, Naturalness, and Accuracy

The results involving users' ratings revealed that learners' judgement of TTS may be affected by the context in which the voices were used. For instance, participants rated TTS comprehensibility, naturalness, and accuracy as equal to the human voice when the task required more than simply emitting an opinion on each category (i.e., understanding a passage to answer a comprehension test and rating the related voices in the text). Participants clearly became more demanding when they were asked to focus exclusively on shorter oral texts (sentences). It was only in this context that they found that TTS sounded less natural or less accurate than human speech samples. These findings corroborate those found in previous research (e.g., Cardoso et al., 2015; Kang et al., 2008; Nusbaum et al., 1995; Stevens et al., 2005).

This difference in judgement may be explained by humans' limited processing capacity. Among several cognitive factors involved in processing a foreign language (e.g., perception, memory), attention plays a fundamental role (Schmidt, 1990). Since attention is a limited cognitive resource that permits participants to focus their mental capacity on individual items (Delogu et al., 1998), cognitively demanding contexts may force attention away from peripheral information (in this case, perceptions of naturalness and accuracy) in order to process the content information conveyed in the speech. In this sense, participants may have shifted their attention to the text content so that they could comprehend the stories, thereby blurring any existing distinctions between TTS and human voices. When the cognitive load was lower, as with the sentence ratings, they attended to those distinctions more clearly and, consequently, they fine-tuned their speech perceptions.

Finally, for the last rating category, the results show that TTS and human voice were judged equally comprehensible for both short stories and sentences. These results do not support the findings reported in Cardoso et al. (2015), who found that the samples produced by the TTS system were rated significantly lower than those that were human-produced. This finding confirms the hypothesis that a change in learning environment (from second to foreign) could positively affect learners' perceptions and attitudes towards TTS-produced input and suggests that *EFL* learners are less sensitive to distinctions between natural and artificial voices than *ESL* students. Low exposure to the target language and the resulting lack of L2 input in the foreign language environment may explain this difference, because when compared to ESL learners, EFL students have fewer opportunities to create strong and more accurate phonological representations of the L2.

Potential for Focus on a Linguistic Feature

The synthetic voice used in this evaluation was also able to match the natural voice in an identification task involving a morpho-phonological feature: The pronunciation of past *-ed*. No difference between voice sources was found in recognizing the presence of past tense morpho-phonology. As such, these findings corroborate those found in Cardoso et al. (2015) regarding the opportunities afforded by TTS voices for students to notice distinctions in L2 input. These results may be explained by Julie's (the TTS voice) accuracy in reproducing English morpho-phonological patterns, as observed in a recent study by John and Cardoso (2016), in which the authors carried out a systematic evaluation of segmental and prosodic features of TTS and human output in order to establish the phonetic accuracy of the synthetic voice. In their evaluation (based on purely phonetic comparisons conducted by the researchers), problematic features of English phonology were targeted, including the TTS's ability to accurately reproduce past *-ed* allomorphy. Their results suggest that TTS performs equally to humans in pronouncing *-ed* forms and, in some contexts (e.g., producing the allomorph /d/), may even surpass humans. Based on our findings, supported by John and Cardoso's research, we may conclude that TTS-generated voices' ability to enhance the input for the noticing of past tense marking is at least similar to that of humans.

Conclusion

This study sought to evaluate the voice of a modern TTS in an English as a foreign language environment

based on a set of assessment measures. It found that TTS-generated samples were comparable to human voice with respect to intelligibility, comprehensibility and ability to provide learners with ability to notice linguistic forms (similar to what human speech is capable of). On the other hand, the participants considered the TTS-based voice less natural and less accurate when compared to the human voice in the context of short sentences.

The results obtained suggest that synthetic voices have the potential to deliver intelligible and comprehensible input, similar to human speech. From a pedagogical standpoint, this is beneficial because their use (preferably using a TTS application) can extend the reach of language classrooms by allowing students to practice on their own time and in their own space; more importantly, TTS may enhance (in both quantity and quality) learners' access to the target language. In sum, the pedagogical use of TTS may provide a level-appropriate, user-controlled solution that produces accurate speech models for pronunciation practice and for the development of language awareness (e.g., to raise students' awareness about the different realizations of the past *-ed* inflection), and thus assist in the acquisition of L2 morpho-phonological patterns.

There were several methodological limitations to the study. First, the small number of participants may prevent more assertive conclusions. Moreover, this study only considered intermediate English proficiency and, accordingly, is not able to determine whether this variable affected the results. Additionally, the high number of comparisons may have decreased statistical power; however, most results would remain unchanged even if an alpha level of .05 for statistical significance had been used (i.e., if the number of comparisons were fewer). We also recognize that, from a pedagogical standpoint, the focus-on-form approach of one of the tasks (in which the participants were asked to determine whether the verb was in the past tense or not) may be difficult to implement in real-life situations if it is used as unguided listening practice. Finally, due to the number of tests carried out during the experiment and the time limitations of a one-time study, this research opted for a reduced quantity of tokens for some tasks (e.g., the past *-ed* feature identification task) so as to not overextend the session time or fatigue the participants.

For future voice quality evaluations, the investigation should consider a larger number of participants from different proficiency levels. It would also be wise to divide the experiment into multiple sections with pauses in between so that the number of tokens may be increased without causing participant fatigue. Future studies should also evaluate CALL software using actual TTS applications for language learning: Would the results be different if the participants had access to all features available for TTS in which they can repeat forms at will and manipulate the input in terms of speed, pitch, or regional accent? Finally, to gather empirical evidence of TTS's potential as a pedagogical tool, one should examine whether its use leads to learning gains (e.g., if its use facilitates the acquisition of regular past tense allomorphy), over an extended period of use.

From a pedagogical perspective, Leow (2015) believes that it is the learners' responsibility to learn (as no one can learn for them) and to come to class prepared to practice, whereas teachers should offer students well-designed tasks to maximize their learning. In this context, TTS may help teachers develop suitable and personalized learning tasks for their students and have the potential to enhance the L2 learning environment by affording students the opportunity to select their own materials and, consequently, have an active role in the learning process.

Acknowledgements

We would like to acknowledge partial funding from the Canadian Social Sciences and Humanities Research Council to Walcir Cardoso and Laura Collins (SSHRC 435-2016-1603), and thank the people who were involved in different aspects of this study, particularly George Smith and Cesar Garcia Fuentes, who helped us conceptualize parts of our research instruments, Jennica Grimshaw, who lent us their amazing voice to represent the human speech, and Randall Halter, for his statistical assistance. Finally, we would like to acknowledge the invaluable contribution of the participants.

References

Bailly, G. (2003). Close shadowing natural versus synthetic speech. *International Journal of Speech Technology, 6*(1), 11–19.

Cardoso, W., Smith, G., & Garcia Fuentes, C. (2015). Evaluating text-to-speech synthesizers. In F. Helm, L. Bradley, M. Guarda, & S. Thouësny (Eds.), *Critical CALL—Proceedings of the 2015 EUROCALL conference*, Padova, Italy (pp. 108–113). Dublin, IE: Researchpublishing.net.

British Council Brasil (2015). *O Ensino de Inglês na Educação Pública Brasileira.* São Paulo, BR: British Council.

Chapelle, C. (2001). *Computer applications in second language acquisition: Foundations for teaching testing and research.* Cambridge, UK: Cambridge University Press.

Chapelle, C. (2003). *English language learning and technology: Lectures on applied linguistics in the age of information and communication.* Amsterdam, NL: John Benjamins.

Collins, L., & Muñoz, C. (2016). The foreign language classroom: Current perspectives and future considerations. *The Modern Language Journal, 100*(1), 133–147.

Collins, L., Trofimovich, P. White, J., Cardoso, W. & Horst, M. (2009). Some input on the easy/difficult grammar question. *Modern Language Journal, 93*(3), 336–353.

Dall, R., Yamagishi, J., & King, S. (2014). Rating naturalness in speech synthesis: The effects of style and expectation. *Proceedings of Speech Prosody*. Retrieved from http://www.cstr.ed.ac.uk/downloads/publications/2014/Dall_Yamagishi_King_SpeechProsody2014.pdf

Delogu, C., Conte, S., & Sementina, C. (1998). Cognitive factors in the evaluation of synthetic speech. *Speech Communication, 24*(2), 153–168.

Derwing, T. & Munro, M. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39*(3), 379–397.

Díez-Bedmar, M. & Pérez-Paredes, P. (2012). The types and effects of peer native speakers' feedback on CMC. *Language Learning & Technology, 16*(1), 62–90. https://www.lltjournal.org/item/2761

Grimshaw, J., & Cardoso, W. (2018). Activate space rats! Fluency development in a mobile game-assisted environment. *Language Learning & Technology*, *22*(3), 159–175. https://www.lltjournal.org/item/3086

Handley, Z. (2009). Is text-to-speech synthesis ready for use in computer-assisted language learning? *Speech Communication, 51*(10), 906–919.

Herrington, R. (2002). Research and Statistical Support: Controlling the false discovery rate in multiple hypothesis testing [PDF]. Retrieved from https://it.unt.edu/sites/default/files/rss-id-false-discovery-rate-hypothesis-testing.pdf

Jamieson, J. & Chapelle, C. A. (2010). Evaluating CALL use across multiple contexts. *System, 38*(3), 357–369.

John, P. & Cardoso, W. (2016). A comparative study of text-to-speech and native speaker output. In J. Demperio, E. Rosales & S. Springer (Eds.), *Proceedings of the meeting on English language teaching* (pp. 78–96). Québec: UQAM Press.

Kang, M., Kashiwagi, H., Treviranus, J., & Kaburagi, M. (2008). Synthetic speech in foreign language learning: An evaluation by learners. *International Journal of Speech Technology, 11*(2), 97–106.

Kirstein, M. (2006). *Universalizing universal design: Applying text-to-speech technology to English language learners' process writing* (Doctoral dissertation). University of Massachusetts, Boston, MA. Retrieved from https://search.proquest.com/openview/fa884a7aa83b7cdf43dc49d92e4ca645/1?pq-origsite=gscholar&cbl=18750&diss=y

Leow, R. (2015). Conclusion: The changing L2 classroom, and where do we go from here? In R.P Leow, (Ed.), *Explicit learning in the L2 classroom: A student-centered approach* (pp. 270–278). New York, US.: Routledge.

Liakin, D., Cardoso, W., & Liakina, N. (2015). Learning L2 pronunciation with a mobile speech recognizer: French /y/. *CALICO Journal, 32*(1), 1–25.

Liakin, D., Cardoso, W., & Liakina, N. (2017). The pedagogical use of mobile speech synthesis: Focus on French liaison. *Computer Assisted Language Learning, 30*(3–4), 348–365.

Major, R. (2007). Identifying a foreign accent in an unfamiliar language. *Studies in Second Language Acquisition, 29*(4), 539–556.

Nusbaum, H., Francis, A., & Henly, A. (1995). Measuring the naturalness of synthetic speech. *International Journal of Speech Technology, 2*(1), 7–19.

Ortega, L. (2013). *Understanding second language acquisition.* Abingdon, UK: Routledge.

Proctor, C., Dalton, B., & Grisham, D. (2007). Scaffolding English language learners and struggling readers in a universal literacy environment with embedded strategy instruction and vocabulary support. *Journal of Literacy Research, 39*(1), 71–9.

Qian, M., Chukharev-Hudilainen, E., & Levis, J. (2018). A system for adaptive high-variability segmental perceptual training: Implementation, effectiveness, transfer. *Language Learning & Technology, 22*(1), 69–96. https://www.lltjournal.org/item/3032

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11,* 129–158.

Soler-Urzua, F. (2011). *The acquisition of English /ɪ/ by Spanish speakers via text-to-speech synthesizers: A quasi-experimental study* (Master's Thesis). Concordia University, Montreal, CA.

Stern, S., Mullennix, J., & Yaroslavsky, I. (2006). Persuasion and social perception of human vs. synthetic voice across person as source and computer as source conditions. *International Journal of Human-Computer* Studies, *64*(1), 43–52.

Stevens, C., Lees, N., Vonwiller, J., & Burnham, D. (2005). On-line experimental methods to evaluate text-to-speech (TTS) synthesis: Effects of voice gender and signal quality on intelligibility, naturalness and preference. *Computer Speech & Language, 19*(2), 129–146.

Tanaka, T. (2009). Communicative language teaching and its cultural appropriateness in Japan. *Doshisha Studies in English, 84*, 107–123

Thomson, R. (2012). Improving L2 listeners' perception of English vowels: A computer–mediated approach. *Language Learning, 62*(4), 1231–1258.

Thomson, R. (2018). High Variability [Pronunciation] Training (HVPT). *Journal of Second Language Pronunciation, 4*(2), 208–231.

About the Authors

Tiago Bione is a PhD student in Applied Linguistics at Concordia University. His research focuses on the use of music and speech technology in second language learning with a focus on pronunciation.

E-mail: tiagobione@gmail.com

Walcir Cardoso is a Professor of Applied Linguistics at Concordia University. He conducts research on the effects of computer technology (e.g., clickers, text-to-speech synthesizers, automatic speech recognition) on L2 learning and the L2 acquisition of phonology, morphosyntax, and vocabulary.

E-mail: walcir.cardoso@concordia.ca